# Estimating Smooth GLM in Non-interactive Local Differential Privacy Model with Public Unlabeled Data

**Di Wang** [*] KAUST
**Huanyu Zhang** Cornell University
**Marco Gaboardi** SUNY at Buffalo
**Jinhui Xu** SUNY at Buffalo

## Abstract

In this paper, we study the problem of estimating smooth Generalized Linear Models (GLM) in the Non-interactive Local Differential Privacy (NLDP) model. Different from its classical setting, our model allows the server to access some additional public but unlabeled data. By using Stein's lemma and its variants, we first show that there is an $(\epsilon, \delta)$-NLDP algorithm for GLM (under some mild assumptions), if each data record is i.i.d sampled from some sub-Gaussian distribution with bounded $\ell_1$-norm. Then with high probability, the sample complexity of the public and private data, for the algorithm to achieve an $\alpha$ estimation error (in $\ell_\infty$-norm), is $O(p^2\alpha^{-2})$ and $O(p^2\alpha^{-2}\epsilon^{-2})$, respectively, if $\alpha$ is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$), where $p$ is the dimensionality of the data. This is a significant improvement over the previously known quasi-polynomial (in $\alpha$) or exponential (in $p$) complexity of GLM with no public data. We then extend our idea to the non-linear regression problem and show a similar phenomenon for it. Finally, we demonstrate the effectiveness of our algorithms through experiments on both synthetic and real world datasets. To our best knowledge, this is the first paper showing the existence of efficient and effective algorithms for GLM and non-linear regression in the NLDP model with public unlabeled data.

## 1 Introduction

Generalized Linear Model (GLM) is one of the most fundamental models in statistics and machine learning. It generalizes ordinary linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. GLM was introduced as a way of unifying various statistical models, including linear, logistic and Poisson regressions. It has a wide range of applications in various domains, such as social sciences [39], genomics research [29], finance [24] and medical research [23]. The model can be formulated as follows.

**GLM:** Let $y \in [0, 1]$ be the response variable that belongs to an exponential family with natural parameter $\eta$. That is, its probability density function can be written as $p(y|\eta) = \exp(\eta y - \Phi(\eta))h(y)$, where $\Phi$ is the *cumulative generating function*. Given observations $y_1, \cdots, y_n$ such that $y_i \sim p(y_i|\eta_i)$ for $\eta = (\eta_1, \cdots, \eta_n)$, the maximum likelihood estimator (MLE) can be written as $p(y_1, y_2, \cdots |\eta) = \exp(\sum_{i=1}^n y_i\eta_i - \Phi(\eta_i))\Pi_{i=1}^n h(y_i)$. In GLM, we assume that $\eta$ is modeled by linear relations, *i.e.,* $\eta_i = \langle x_i, w^*\rangle$ for some $w^* \in \mathbb{R}^p$ and feature vector $x_i$. Thus, maximizing MLE is equivalent to minimizing $\frac{1}{n}\sum_{i=1}^n[\Phi(\langle x_i, w\rangle) - y_i\langle x_i, w\rangle]$. The goal is to find $w^*$, which is equivalent to

[*]The first two authors contribute equally. Part of the work was done when D.W and M.G were visiting Simons Institute for the Theory of Computing.

Preprint. Under review.

minimizing its population version

$$w^* = \arg\min_{w \in \mathbb{R}^p} \mathbb{E}_{(x,y)}[\Phi(\langle x, w\rangle) - y\langle x, w\rangle]. \tag{1}$$

One often encountered challenge for using GLM in real world applications is how to handle sensitive data, such as those in social science and medical research. As a commonly-accepted approach for preserving privacy, Differential Privacy (DP) [15] provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

As a popular way of achieving DP, Local Differential Privacy (LDP) has received considerable attentions in recent years and been adopted in industry [12, 17, 30]. In LDP, each individual manages his/her proper data and discloses them to a server through some DP mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. Information exchange between the server and each individual could be either only once or multiple times. Correspondingly, protocols for LDP are called non-interactive LDP (NLDP) or interactive LDP. Due to its ease of implementation (*e.g.* no need to deal with the network latency problem), NLDP is often preferred in practice.

While there are many results on GLM in the DP and interactive LDP models [9, 4, 20, 21], GLM in NLDP is still not well understood due to the limitation of the model. [27, 35, 40] and [36] comprehensively studied this problem. However, all of these results are on the negative side. More specifically, they showed that to achieve an error of $\alpha$, the sample complexity needs to be quasi-polynomial in $\alpha$ [36, 40] or even exponential in the dimensionality $p$ [27, 35] (see Related Work section for more details). Due to these negative results, there is no study on the practical performance of these algorithms.

To address this high sample complexity issue of NLDP, a possible way is to make use of some recent developments on the central DP model. Quite a few results [3, 19, 25, 26, 5] have suggested that by allowing the server to access some public but unlabeled data in addition to the private data, it is possible to reduce the sample complexity in the central DP model, under the assumption that these public data have the same marginal distribution as the private ones. It has also shown that such a relaxed setting is likely to enable better practical performance for problems like Empirical Risk Minimization (ERM) [19, 25]. Thus, it would be interesting to know whether the relaxed setting can also help reduce sample complexity in the NLDP model.

With this thinking, our main questions now become the follows. **Can we further reduce the sample complexity of GLM in the NLDP model if the server has additional public but unlabeled data? Moreover, is there any efficient algorithm for this problem in the relaxed setting?**

In this paper, we provide positive answers to the above two questions. Our contributions can be summarized as follows:

1. We first show that when the feature vector $x$ of GLM is sub-Gaussian with bounded $\ell_1$-norm, there is an $(\epsilon, \delta)$-NLDP algorithm for GLM (under some mild assumptions) whose sample complexities of the private and public data, for achieving an error of $\alpha$ (in $\ell_\infty$-norm), are $O(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$ (with other terms omitted), respectively, if $\alpha$ is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$). We note that this is the first result that achieves a **fully polynomial** sample complexity for a general class of loss functions in the NLDP model with public unlabeled data.
2. We then extend our idea to the non-linear regression problem. By using the zero-bias transformation [18], we show that when $x$ is sub-Gaussian with bounded $\ell_1$-norm, it exhibits the same phenomenon as GLM.
3. Finally, we provide an experimental study of our algorithms on both synthetic and real world datasets. The experimental results suggest that our methods are efficient and effective, which is consistent with our theoretical analysis. To our best knowledge, these are the **first** effective algorithms in the NLDP model with public unlabeled data for both the GLM and non-linear regression problems.

Due to space limit, some background, lemmas, all the proofs and additional experimets are included in the Appendix of supplementary material. Moreover, the codes of experimental are also included in supplementary material.

## 2 Related Work

Private learning with public unlabeled data has been studied previously in [19, 25, 26, 5]. These results differ from ours in quite a few ways. Firstly, all of them consider either the multiparty setting or the centralized model. Consequently, none of them can be used to solve our problems. Specifically, [19] considered the multiparty setting where each party possesses several data records, while each party in our NLDP model has only one data record. [25, 26] investigated the DP model, used sub-sample and aggregate to train some deep learning models, but provided no provable sample complexity. [5] also studied the DP model by combining the distance to instability and the sparse vector techniques, and showed some theoretical guarantees. However, both the sub-sample/aggregate and the sparse vector methods cannot be used in the NLDP model. Moreover, public data in their methods are also used quite differently from ours. Secondly, all of the above results use the private data to label the public data and conduct the learning process on the public data, while we use the public data to approximate some crucial constants. Finally, all of the previous methods rely on the known model or loss functions, while in our algorithms the loss functions could be unknown to the users; also the server could use multiple loss functions with the same sample complexity.

The problems considered in this paper can be viewed as restricted versions of the general ERM problem in the NLDP model. Due to its challenging nature, ERM in NLDP has only been considered in a few papers, such as [27, 35, 36, 40, 10, 37]. [27] gave the first result on convex ERM in NLDP and provided an algorithm with a sample complexity of $O(2^p\alpha^{-(p+1)}\epsilon^{-2})$. They showed that the exponential dependency on the dimensionality $p$ is not avoidable for general loss functions when restricting on some oracles. Later, [35] showed that when the loss function is smooth enough, the exponential term of $\alpha^{-\Omega(p)}$ can be reduced to polynomial, but not the other exponential terms. Recently, [36] further showed that the sample complexity for any 1-Lipschitz convex GLM can be reduced to linear in $p$ and quasi-polynomial in $\alpha$, which extends the work in [40]. In this paper, we show, for the first time, that the sample complexity of GLM can be reduced to fully polynomial with the help of some public but unlabeled data and some mild assumptions on GLM. There are also works on some special loss functions. For example, [37] studied the high dimensional sparse linear regression problem and [10] considered the problem of learning halfspaces with polynomial samples. Since these results are only for some special loss functions (instead of a family of functions), they are incomparable with ours.

## 3 Preliminaries

**Local Differential Privacy (LDP):** In LDP, we have data universe $\mathcal{X}$ and $\mathcal{Y}$, $n$ players with each holding a private data record $(x, y) \in \mathcal{X} \times \mathcal{Y}$ sampled from some distribution $\mathcal{P}$, where $x \in \mathbb{R}^p$ is the feature vector and $y \in \mathbb{R}$ is the label or response, and a server that is in charge of coordinating the protocol. An LDP protocol proceeds in $T$ rounds. In each round, the server sends a message, which is often called a query, to a subset of the players, requesting them to run a particular algorithm. Based on the query, each player $i$ in the subset selects an algorithm $Q_i$, runs it on her own data, and sends the output back to the server.

**Definition 1** ([22])**.** An algorithm $Q$ is $\epsilon$-locally differentially private (LDP) if for all pairs $x, x' \in \mathcal{D}$, and for all events $E$ in the output space of $Q$, we have $\text{Pr}[Q(x) \in E] \leq e^{\epsilon}\text{Pr}[Q(x') \in E]$. A multi-player protocol is $\epsilon$-LDP if for all possible inputs and runs of the protocol, the transcript of player i's interaction with the server is $\epsilon$-LDP. If $T = 1$, we say that the protocol is $\epsilon$ **non-interactive LDP (NLDP)**.

**Our Model:** Different from the above classical NLDP model where only one private dataset $\{(x_i, y_i)\}_{i=1}^n$ exists, the NLDP model in our setting allows the server to have an additional public but unlabeled dataset $D' = \{x_j\}_{j=n+1}^{n+m} \subset \mathcal{X}^m$, where each $x_j$ is sampled from $\mathcal{P}_x$, which is the marginal distribution of $\mathcal{P}$ (*i.e.,* they have the same distribution as $\{x_i\}_{i=1}^n$).

## 4 Privately Estimating Generalized Linear Models

In this section, we study GLM in our model and privately estimate $w^*$ in (1) by using both the private data $\{(x_i, y_i)\}_{i=1}^n$ and the public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$. Our goal is to achieve a fully polynomial sample complexity for $n$ and $m$, *i.e.,* $n, m = \text{Poly}(p, \frac{1}{\epsilon}, \frac{1}{\alpha}, \log\frac{1}{\delta})$, such that there is an $(\epsilon, \delta)$-NLDP algorithm with estimation error less than $\alpha$ (with high probability). Before presenting our ideas, we first consider the following lemma for $x \sim \mathcal{N}(0, \Sigma)$, which is from Stein's lemma [8].

**Algorithm 1** Non-interactive LDP for smooth GLM with public data

**Input**: Private data $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{R}^p \times \{0,1\})^n$, where $\|x_i\|_1 \leq r$ and $|y_i| \leq 1$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.

**for** Each user $i \in [n]$ **do**
- Release $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log\frac{2.5}{\delta}}{\epsilon^2})$.
- Release $\widehat{x_i y_i} = x_i y_i + E_{2,i}$, where $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2 \log\frac{2.5}{\delta}}{\epsilon^2} I_p)$.

**end for**

**for** The server **do**
- Let $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1}\widehat{X^T y}$.
- Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$. Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{m}\sum_{j=n+1}^{n+m} \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$ by using Newton's root-finding method (or other methods):
- **for** $t = 1, 2, \cdots$ until convergence **do**
  - $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1}\Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j \Phi^{(3)}(c\tilde{y}_j)\}}$.
- **end for**

**end for**

Return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.

**Lemma 1** ([8])**.** If $x \sim \mathcal{N}(0, \Sigma)$, then $w^*$ in (1) can be written as $w^* = c_\Phi \times w^{ols}$, where $c_\Phi$ is the fixed point of $z \mapsto (\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)])^{-1}$ (if we assume taht $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)] \neq 0$) and $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$ is the Ordinary Least Squares (OLS) vector. [2]

From Lemma 1, we can see that to obtain $w^*$, it is sufficient to estimate $w^{ols}$ and the underlying constant $c_\Phi$. Specifically, to estimate $w^{ols}$ in a non-interactive local differentially private manner, a direct way is to let each player perturb her sufficient statistics, *i.e.,* $x_i x_i^T$ and $y_i x_i$. After receiving the private OLS estimator $\hat{w}^{ols}$, the server can then estimate the constant $c_\Phi$ by using the public unlabeled data and $\hat{w}^{ols}$. From the definition, it is easy to see that $c_\Phi$ is independent of the label $y$. Thus, $c_\Phi$ can be estimated by using the empirical version of $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)]$. That is, find the root of the function $1 - \frac{c}{m}\sum_{j=n+1}^{n+m}\Phi^{(2)}(c\langle x_j, \hat{w}^{ols}\rangle)$. Several methods are available for finding roots, and in Algorithm 1 we use the Newton's method which has a quadratic convergence rate.

One problem with the above approach is that Lemma 1 needs $x$ to be Gaussian, which implies that the sensitivity of the term $x_i x_i^T$ could be unbounded. We also note that Lemma 1 is only for Gaussian distribution. The following lemma extends Lemma 1 to bounded sub-Gaussian with an additional additive error of $O(\frac{\|w^*\|_\infty^2}{\sqrt{p}})$.

**Lemma 2** ([16])**.** Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector $x$ that is sub-Gaussian with zero mean, whose covariance matrix $\Sigma$ has its corresponding $\Sigma^{\frac{1}{2}}$ being diagonally dominant [3], and whose distribution is supported on a $\ell_2$-norm ball of radius $r$. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of $x$ with sub-Gaussian norm $\|v\|_{\psi_2} = \kappa_x$. If each $v_i$ has constant first and second conditional moments (*i.e.,* $\forall j \in [p]$ and $\tilde{w} = \Sigma^{\frac{1}{2}}w^*$, $\mathbb{E}[v_{ij}|\sum_{k\neq j}\tilde{w}v_{ik}]$ and $\mathbb{E}[v_{ij}^2|\sum_{k\neq j}\tilde{w}v_{ik}]$ are deterministic) and the function $\Phi^{(2)}$ is Lipschitz continuous with constant $G$, then for $c_\Phi = \frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)]}$ (assuming $\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)] \neq 0$), the following holds for GLM in (1)

$$\|\frac{1}{c_\Phi}\cdot w^* - w^{ols}\|_\infty \leq O(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty^2}{\sqrt{p}}), \tag{2}$$

[2] $\Phi^{(2)}$ is the second order derivative of function $\Phi$, the similar to $\Phi^{(3)}$.

[3] A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row.

where $\rho_q$ for $q = \{2, \infty\}$ is the conditional number of $\Sigma$ in $\ell_q$ norm and $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$ is the OLS vector.

Lemma 2 indicates that we can use the same idea as above to estimate $w^*$. Note that the forms of $c_\Phi$ in Lemmas 1 and 2 are different. However, due to the closeness of $w^*$ and $w^{ols}$ in (2), we can still use $\frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle \bar{c}_\Phi])}$ to approximate $c_\Phi$, where $\bar{c}_\Phi$ is the root of $c\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle c)] - 1$ and it could be approximated by using the public unlabeled data. Combining these ideas, we have Algorithm 1.

**Theorem 1.** For any $0 < \epsilon, \delta < 1$, Algorithm 1 is $(\epsilon, \delta)$ non-interactive LDP.

The following theorem shows the sample complexity of the bounded sub-Gaussian case.

**Theorem 2.** Under the assumptions of Lemma 2, if further assume that the distribution of $x$ is supported on the $\ell_1$-norm ball with radius $r$, $|\Phi^{(2)}(\cdot)| \leq L$, and for some constant $\bar{c}$ and $\tau > 0$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ satisfies the condition of $f(\bar{c}) \geq 1 + \tau$, and the derivative of $f$ in the interval $[0, \max\{\bar{c}, c_\Phi\}]$ does not change the sign (*i.e.,* its absolute value is lower bounded by some constant $M > 0$), then for sufficiently large $m, n$ such that

$$m \geq \Omega\big(\|\Sigma\|_2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\rho_2\rho_\infty^2 p^2\big), n \geq \Omega\big(\frac{\rho_2\rho_\infty^2\|\Sigma\|_2^2 p^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big), \tag{3}$$

with probability at least $1 - \exp(-\Omega(p)) - \xi$, the output $\hat{w}^{glm}$ in Algorithm 1 satisfies

$$\|\hat{w}^{glm} - w^*\|_\infty \leq O\big(\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}}$$
$$+ \frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + \frac{\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\big), \tag{4}$$

where $G, L, \tau, M, \bar{c}, r, \kappa_x, \frac{1}{c_\Phi}$ are assumed to be $O(1)$ and thus omitted in the Big-$O$ notations (see Appendix for the explicit form of $m$ and $n$).

**Corollary 1.** Theorem 2 suggests that if we omit all the other terms and assume that $\|w^*\|_\infty = O(1)$, then for any given error $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$, there is an $(\epsilon, \delta)$-LDP algorithm whose sample complexity of private ($n$) and public unlabeled ($m$) data, to achieve an estimation error of $\alpha$ (in $\ell_\infty$-norm), is $O(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$, respectively. We note that $m \leq n$, which means that the sample complexity of the public data is less than that of the private data. We also note that the sample complexity of the public data is independent of the privacy parameters $\epsilon, \delta$. All these are quite reasonable in practice. We will also see that in practice we do not need large amount of public data (see Section E.1 in Appendix for details).

There are also some previous work on LDP linear regression. [27] proposed an algorithm with a sample complexity of $\tilde{O}(p\alpha^{-2}\epsilon^{-2})$ and [40] achieved a sample complexity of $O(\log p\alpha^{-4}\epsilon^{-2})$. It seems that our sample complexity for the more general GLM is worse than theirs. However, these results are not really comparable due to their different settings. Firstly, [27, 40] considered the optimization error and [37] measured the $\ell_2$-norm statistical error, while we estimate the $\ell_\infty$-norm statistical error. Secondly, $w^*$ is assumed to be bounded in $\ell_2$-norm in [27], $\ell_1$-norm in [40], and $\ell_\infty$-norm in ours. Finally, in this paper, we have some additional assumptions on the data distribution compared with the previous results. There is also a result on NLDP linear regression [37]. It relies on assumptions that $\|x\|_2 = O(\sqrt{p})$ and $w^*$ is 1-sparse, which are not needed in ours.

Also note that in Theorem 2, $\Phi^{(2)}$ is assumed to be bounded. This is a quite common assumption in related works such as [35, 34]. Actually, this condition can be relaxed by only assuming that $\Phi^{(2)}(\langle x, w\rangle)$ is sub-Gaussian in some range of $w$.

**Theorem 3.** Under the assumptions of Lemma 2, if further assume that the distribution of $x$ is supported on the $\ell_1$-norm ball with radius $r$, $\sup_{w:\|w-\Sigma^{\frac{1}{2}}w^{ols}\|_2\leq 1}\|\Phi^{(2)}(\langle x, w\rangle)\|_{\psi_2} \leq \kappa_g$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ satisfies the inequality of $f(\bar{c}) \geq 1 + \tau$ for some constant $\bar{c}$ and $\tau > 0$, and the derivative of $f$ in the interval of $[0, \max\{\bar{c}, c_\Phi\}]$ does not change the sign (*i.e.,* its absolute value is lower bounded by some constant $M > 0$), then for sufficiently large $m, n$ such that

$$m \geq \tilde{\Omega}\big(\frac{1}{\tilde{\mu}^2}\epsilon^2 n\big), n \geq \Omega\big(\frac{p^2\rho_2\rho_\infty^2\|\Sigma\|_2^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big), \tag{5}$$

the following holds with probability at least $1 - \exp(-\Omega(p)) - \xi$,

$$\|\hat{w}^{glm} - w^*\|_\infty \leq O\Big(\frac{p\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\|w^*\|_\infty \max\{1, \|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} + \frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}}{\sqrt{p}} + \sqrt{\rho_2}\rho_\infty\|w^*\|_\infty\max\{1, \|w^*\|_\infty\}\frac{1}{\tilde{\mu}}\sqrt{\frac{p^2\log m}{m}}\Big), \tag{6}$$

where $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$, the terms of $r, \kappa_x, \kappa_g, G, M, \tau, \bar{c}, \frac{1}{c_\Phi}$ are assumed to be constants, and thus omitted in the Big-$O$ notations (see Appendix for the explicit forms of $m$ and $n$).

From the above theorem, we can see that with more relaxed assumptions, the sample complexity in Theorem 3 increases by a factor of $O(\log m)$ to achieve an upper bound on the statistical error ( in $\ell_\infty$-norm) that is asymptotically the same as the one in Theorem 2.

**Remark 1.** Algorithm 1 has several advantages over existing techniques. Firstly, different from the approach of using Gradient Descent methods to solve DP-ERM (*e.g.,* [38]), our algorithm is parameter-free. That is, we do not need to choose a specific step size, an iteration number or initial vectors. Secondly, comparing with some previous work such as [40, 27, 36], all of our above results do not need to assume that the loss function is convex. Thirdly, since the private data contributes only to obtaining the OLS estimator, and only the constant $\hat{c}_\Phi$ depends on the loss function $\Phi$, this means that with probability at least $1 - T\exp(-\Omega(p)) - \xi$, our algorithm can simultaneously use $T$ different loss functions to achieve the same errors and with the same sample complexity. This implies that we can answer at most $O(\exp(O(p))$ number of GLM queries with constant probability to achieve error $\alpha$ for each query with the same sample complexity as in Theorem 2. To our best knowledge, this is the first result which can answer multiple non-linear queries in the NLDP model with polynomial sample complexity. Previous results are either for linear queries [7, 2], or in the central DP model [32].

A not so desirable issue of Theorems 2 and 3 is that they need quite a few assumptions/conditions. Although almost all of them commonly appear in some related work, the assumptions on function $f$ seem to be a little weird. They are introduced to ensure that the function $f - 1$ has a root and $\hat{c}_\Phi$ is close to $c_\Phi$ for large enough $m$. Fortunately, this is a not big issue in practice. As shown in [16], these conditions actually hold for many loss functions, such as logistic and boosting loss. Also, as we will see later, our experiments show that the algorithm actually performs quite well for many loss functions that may not satisfy these assumptions. Also, we note that the error bounds in Theorems 2 and 3 are dependent on the $\ell_1$-norm of the upper bound of $x_i$, while such a dependency is on the $\ell_2$-norm in previous work such as [27, 40]. We leave the problem of relaxing/lifting these assumptions to future research. It is also notable that public unlabeled data is only used in Step 6-10 in Algorithm 1, where we use it to find a root of some function. It is still an open problem whether we can extend to the case where the serve only hold private unlabeled data.

## 5 Privately Estimating Non-linear Regressions

In this section, we extend our ideas in the previous section to the problem of estimating non-linear regression in the NLDP model with public unlabeled data. We assume that there is an underlying vector $w^* \in \mathbb{R}^p$ with $\|w^*\|_2 \leq 1$ such that

$$y = f(\langle x, w^*\rangle) + \sigma, \tag{7}$$

where $x$ is the feature vector sampled from some distribution (for simplicity, we assume that the mean is zero) and $y$ is the response. $\sigma$ is the zero-mean noise which is independent of $x$ and bounded by some constant $C = O(1)$ (*i.e.,* $\sigma \in [-C, C]$). $f$ is some known differentiable link function with $f(0) \neq \infty$ [4]. We note that these assumptions are quite common in related work such as [37, 14]. In our model, the goal is to obtain some estimator $w^{\text{priv}}$ of $w^*$, based on the private dataset $\{(x_i, y_i)\}_{i=1}^n$ and the public unlabeled dataset $\{x_j\}_{j=n+1}^{n+m+1}$ via some NLDP algorithms.

To solve this problem, we first use the zero-bias transformation [18] and the techniques in [16] to get a lemma similar to Lemma 2.

[4] This assumption can be relaxed to "there is a point $x$ such that $f(x) \neq 0$".

**Definition 2** (Zero-bias Transformation)**.** Let $z$ be a random variable with mean 0 and variance $\sigma^2$. Then, there exists a random variable $z^*$ that satisfies $\mathbb{E}[zf(z)] = \sigma^2\mathbb{E}[f'(z^*)]$ for all differentiable functions $f$. The distribution of $z^*$ is called the $z$-zero-bias distribution.

**Theorem 4.** Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be $n$ i.i.d realizations of a random vector $x$ which is sub-Gaussian with zero mean, whose covariance matrix $\Sigma$ has its $\Sigma^{\frac{1}{2}}$ being diagonally dominant, and whose distribution is supported on an $\ell_2$-norm ball of radius $r$. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of $x$ with sub-Gaussian norm $\|v\|_{\psi_2} = \kappa_x$. If each $v_i$ has constant first and second conditional moments and function $f'$ is Lipschitz continuous with constant G, then for $c_f = \frac{1}{\mathbb{E}[f'(\langle x_i, w^*\rangle)]}$, the following holds, where $w^{ols}$ is the OLS vector.

$$\|\frac{1}{c_f}\cdot w^* - w^{ols}\|_\infty \leq O(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty^2}{\sqrt{p}}).$$

---

**Algorithm 2** Non-interactive LDP for smooth Non-linear Regression with public data

---

**Input**: Private data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \{0,1\}$ with $\|x_i\|_1 \leq r$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$. Link function $f: \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.

1: **for** Each user $i \in [n]$ **do**

2: Release $\widehat{x_ix_i^T} = x_ix_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4\log\frac{2.5}{\delta}}{\epsilon^2})$. Release $\widehat{x_iy_i} = x_iy_i + E_{2,i}$, where the vector $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2(Lr+|f(0)|+C)^2\log\frac{2.5}{\delta}}{\epsilon^2}I_p)$.

3: **end for**

4: **for** The server **do**

5: Denote $\widehat{X^TX} = \sum_{i=1}^n \widehat{x_ix_i^T}$ and $\widehat{X^Ty} = \sum_{i=1}^n \widehat{x_iy_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^TX})^{-1}\widehat{X^Ty}$.

6: Calculate $\tilde{y}_j = x_j^T\hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$. Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{m}\sum_{j=n+1}^{n+m} f'(\hat{c}_\Phi\tilde{y}_j)$ using Newton's root finding method:

7: **for** $t = 1, 2, \cdots$ until convergence **do**

8: $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1} f'(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{f'(c\tilde{y}_j) + c\tilde{y}_jf^{(2)}(c\tilde{y}_j)\}}$.

9: **end for**

10: **end for**

11: Return $\hat{w}^{nlr} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.

---

From Theorem 4, we can see that it shares the same phenomenon as Lemma 2 (*i.e.,* the OLS vector with some constant could approximate $w^*$ well). Thus, a similar idea to Algorithm 1 can be used to solve this problem for the bounded sub-Gaussian case, which gives us Algorithm 2 and the following theorem.

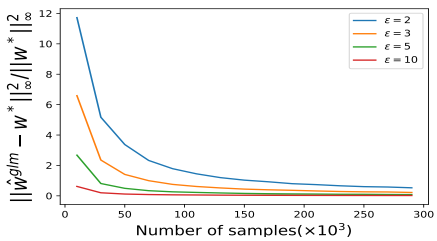
**Theorem 5.** Under the assumptions of Theorem 4, if further assume that the assumptions in Theorem 2 hold for function $f'(\cdot)$ instead of $\Phi^{(2)}(\cdot)$, then for sufficiently large $m, n$ such that

$$m \geq \Omega\big(\|\Sigma\|_2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\rho_2\rho_\infty^2p^2\big), n \geq \Omega\big(\frac{\rho_2\rho_\infty^2\|\Sigma\|_2^2p^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)}\big), \quad (8)$$

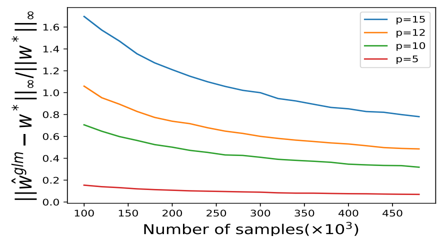
with probability at least $1 - \exp(-\Omega(p)) - \xi$, the output of Algorithm 2 satisfies

$$\|\hat{w}^{nlr} - w^*\|_\infty \leq O\big(\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}} + \frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + \frac{\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\|w^*\|_\infty^3\max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\big), \quad (9)$$

where the terms of $G, L, \tau, M, \bar{c}, r, \kappa_x, C, \frac{1}{c_f}$ are assumed to be $O(1)$ and thus omitted in the Big-$O$ notations (see Appendix for the explicit form of $m$ and $n$).




(a)




(b)

Figure 1: GLM with logistic loss under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality.

# 6 Experiments

## 6.1 Evaluation on synthetic data

**Experimental Setting** For GLM, we consider the problem of binary logistic loss *i.e.,* $\Phi(\langle x, w\rangle) = \ln(1 + \exp(\langle x, w\rangle))$ in (1) while for non-linear regression we set $f(x) = \frac{1}{3}x^3$ in (7). For each problem we first compare the squared relative error $\frac{\|\hat{w}-w^*\|_\infty^2}{\|w^*\|_\infty^2}$ with respect to different privacy parameters $\epsilon \in \{10, 5, 3, 2\}$ with $\delta = \frac{1}{n}$. In these experiments, we estimate the squared relative error with the fixed dimensionality $p = 10$ and the population parameter $w^* = (1, 1, ..., 1)/\sqrt{p}$. The sample size $n$ is chosen from the set $10^4 \cdot \{1, 3, 5, ..., 29\}$. We assume that the same amount of public unlabeled data is available. The features are generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$ and the label is generated according to the logistic model or the model (7). In non-linear regression model, $\sigma$ is bounded by $C = 0.001$. The results are shown in Figure 1a and 2a. For each problem we then evaluate the impact of the dimensionality. In these experiments, we fix the privacy parameters $\epsilon = 10^5$, $\delta = \frac{1}{n}$, and tune the dimensionality $p \in \{5, 10, 12, 15\}$. $w^*$s are the same as above. The sample size takes values from $n \in 10^4 \cdot \{10, 12, 14, ..., 48\}$ and the same amount of public unlabeled data is assumed. The responses are generated as the same as above. We measure the performance directly by the relative error. For each experiments above, we run 1000 times and take the average of the errors. The results are shown in Figure 1b and 2b.

From Figure 1a and 2a, we can see that the square of relative error is inversely proportional to the number of samples $n$. In other words, in order to achieve relative error $\alpha$, we only need the number of private samples $n \sim \frac{1}{\alpha^2}$ if we omit the dependency on the other parameters. Besides, we also observe that the square of relative error is proportional to $\frac{1}{\epsilon^2}$, which matches our theoretical result.

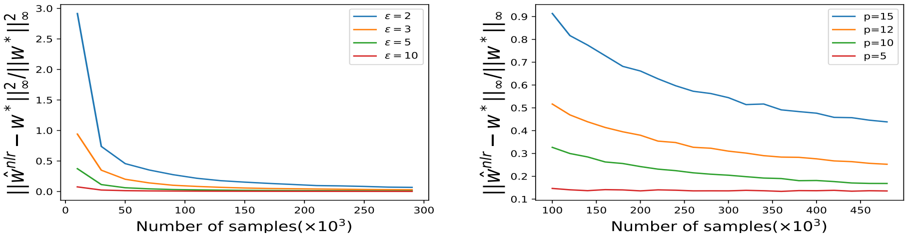
From Figure 1b and 2b, we can see that the relative error increases as the dimensionality increases. It may seem a little weird that it is not linear in the dimensionality. We note that as the dimensionality $p$ changes, some other parameters, for example, the $l_2$ norm of the covariance matrix and $w^*_\infty$ also change, which bring other effects to the relative error.

## 6.2 Evaluation on real data

We conduct experiment for GLM with logistic loss on the Covertype dataset [13] and the SUSY dataset [1] (see Section E in Appendix for details). We measure the performance by the prediction accuracy. For each combination of $\epsilon$ and $n$, the experiment is repeated 1000 times. From Figure 3 we can observe that when $\epsilon$ takes a reasonable value, the performance is approaching to the non-private case, provided that the size of private dataset is large enough. Thus, our algorithm is practical and is comparable to the non-private one.

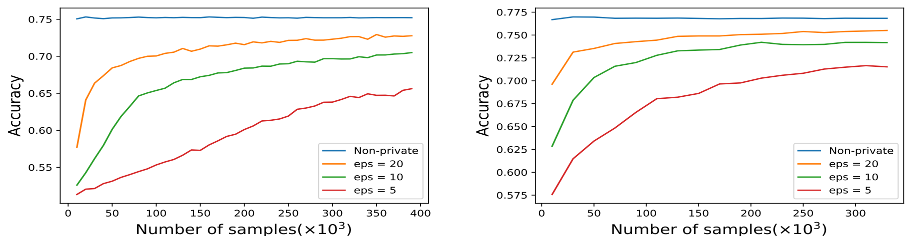
We also study the effect of public unlabeled data, see Section E.1 in Appendix for details.

[5]Note that in the studies on LDP ERM, $\epsilon$ is always chosen as a large value such as [6].




(a) (b)

Figure 2: Cubic regression with i.i.d Bernoulli design. The left plot shows the squared relative error under different level of privacy. The right one shows relative error under different dimensionality.




(a) (b)

Figure 3: GLM with logistic loss on real dataset. Left is for Covertype and right is for SUSY.

## Broader Impact Statement

Privacy is an important ethical issue in the Machine Learning community, in this paper, we mitigate this issue bu studying how to estimate GLM under privacy constraint in a distributed setting. We do not figure out any *substantial* ethic issue regarding this work. However, it might lead to several societal consequences. NLDP model has been adopted in industry, however, there is no practical and effective algorithm on estimating GLM in NLDP model. This is the first paper showing the existence of efficient and effective algorithms for GLM and non-linear regression in the NLDP model with public unlabeled data.

## References


[1] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

[2] Raef Bassily. Linear queries estimation with local differential privacy. *arXiv preprint arXiv:1810.02810*, 2018.

[3] Raef Bassily and Anupama Nandi. Privately answering classification queries in the agnostic pac model. *arXiv preprint arXiv:1907.13553*, 2019.

[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[5] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems*, pages 7102–7112, 2018.

[6] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[7] Jaroslaw Blasiok, Mark Bun, Aleksandar Nikolov, and Thomas Steinke. Towards instance-optimal private query release. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium*

*on Discrete Algorithms*, pages 2480–2497. Society for Industrial and Applied Mathematics, 2019.

[8] David R Brillinger. A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

[9] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[10] Amit Daniely and Vitaly Feldman. Learning without interaction requires separation. *arXiv preprint arXiv:1809.09165*, 2018.

[11] Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, pages 360–368, 2013.

[12] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[14] John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *arXiv preprint arXiv:1806.05756*, 2018.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[16] Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker. Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research*, 20(1):231–275, 2019.

[17] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[18] Larry Goldstein, Gesine Reinert, et al. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4):935–952, 1997.

[19] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563, 2016.

[20] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484, 2014.

[21] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497, 2016.

[22] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[23] James K Lindsey and Bradley Jones. Choosing among generalized linear models applied to medical data. *Statistics in medicine*, 17(1):59–68, 1998.

[24] Alexander J McNeil and Jonathan P Wendin. Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, 14(2):131–149, 2007.

[25] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[26] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[27] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

[28] G. W. Stewart. Matrix perturbation theory, 1990.

[29] Yasuaki Takada, Ryutaro Miyagi, Aya Takahashi, Toshinori Endo, and Naoki Osada. A generalized linear model for decomposing cis-regulatory, parent-of-origin, and maternal effects on allele-specific gene expression. *G3: Genes, Genomes, Genetics*, 7(7):2227–2234, 2017.

[30] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

[31] Terence Tao. Topics in random matrix theory. *Graduate Studies in Mathematics*, 132, 2011.

[32] Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 303–312. ACM, 2015.

[33] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[34] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019.

[35] Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 965–974, 2018.

[36] Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pages 897–902, 2019.

[37] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019.

[38] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

[39] Russell T. Warne. *Statistics for the Social Sciences: A General Linear Model Approach*. Cambridge University Press, 2017.

[40] Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at once, use effectively: Making non-interactive locally private learning possible. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4130–4139. JMLR. org, 2017.

## A Background and Auxiliary Lemmas

**Notations** For a positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$, we define the $M$-norm for a vector $w$ as $\|w\|_M^2 = w^T M w$. $\lambda_{\min}(A)$ is the minimal singular value of the matrix $A$. For a semi positive definite matrix $M \in \mathbb{R}^{p\times p}$, let its SVD composition be $\Sigma = U^T \Sigma U$, where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$, then $M^{\frac{1}{2}}$ is defined as $M^{\frac{1}{2}} = U^T \Sigma^{\frac{1}{2}} U$, where $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

**Definition 3** (Sub-Gaussian)**.** For a given constant $\kappa$, a random variable $x \in \mathbb{R}$ is said to be sub-Gaussian if it satisfies $\sup_{m\geq 1} \frac{1}{\sqrt{m}} \mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$. The smallest such $\kappa$ is the **sub-Gaussian norm** of $x$ and it is denoted by $\|x\|_{\psi_2}$. A random vector $x \in \mathbb{R}^p$ is called a sub-Gaussian vector if there exists a constant $\kappa$ such that for any unit vector $v$, we have $\|\langle x, v\rangle\|_{\psi_2} \leq \kappa$.

**Lemma 3** (Weyl's Inequality [28])**.** Let $X, Y \in \mathbb{R}^{p\times p}$ be two symmetric matrices, and $E = X - Y$. Then, for all $i = 1, \cdots, p$, we have

$$|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2.$$

**Lemma 4.** Let $w \in \mathbb{R}^p$ be a fixed vector and $E$ be a symmetric Gaussian random matrix where the upper triangle entries are i.i.d Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then, with probability at least $1 - \xi$, the following holds for a fixed positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$

$$\|Ew\|_M^2 \leq \sigma^2 \text{Tr}(M) \|w\|^2 \log \frac{2p^2}{\xi}.$$

*Proof of Lemma 4.* Let $M = U^T \Sigma U$ denote the eigenvalue decomposition of $M$. Then, we have

$$\|Ew\|_M^2 = w^T E^T U^T \Sigma U E w = \sum_{i=1}^{p} \sigma_i \sum_{j=1}^{p} [UE]_{ij}^2 w_i^2.$$

Note that $[UE]_{i,j} = \sum_{k=1}^{p} U_{i,k} E_{j,k}$ where $E_{i,j}$ is Gaussian. Since $U$ is orthogonal, we know that $[UE]_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Using the Gaussian tail bound for all $i, j \in [d]^2$, we have

$$\mathbb{P}(\max_{i,j\in[p]^2} |[UE]_{i,j} \geq \sqrt{\sigma^2 \log \frac{2p^2}{\xi}}) \leq \xi.$$

□

**Lemma 5** (Theorem 4.7.1 in [33] )**.** Let $x$ be a random vector in $\mathbb{R}^p$ that is sub-Gaussian with covariance matrix $\Sigma$ and $\|\Sigma^{-\frac{1}{2}} x\|_{\psi_2} \leq \kappa_x$. Then, with probability at least $1 - \exp(-p)$, the empirical covariance matrix $\frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$ satisfies

$$\|\frac{1}{n} X^T X - \Sigma\|_2 \leq C \kappa_x^2 \sqrt{\frac{p}{n}} \|\Sigma\|_2.$$

**Lemma 6** (Corollary 2.3.6 in [31])**.** Let $M \in \mathbb{R}^{p\times p}$ be a symmetric matrix whose entries $m_{ij}$ are independent for $j > i$, have mean zero, and are uniformly bounded in magnitude by 1. Then, there exists absolute constants $C_2, c_1 > 0$ such that with probability at least $1 - \exp(-C_2 c_, 1p)$, the following inequality holds $\|M\|_2 \leq C\sqrt{p}$.

Below we introduce some concentration lemmas given in [16].

**Lemma 7.** Let $\mathbb{B}^\delta(\tilde{w})$ denote the ball centered at $\tilde{w}$ and with radius $\delta$ (*i.e.*, $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2\cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d isotropic sub-Gaussian random vectors with $\|x_i\|_{\psi_2} \leq k_x$, and $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is Lipschitz continuous with $G$ and satisfies $\sup_{w\in\mathbb{B}^\delta(\tilde{w})} \|g(\langle x, w\rangle)\|_{\psi_2} \leq \kappa_g$, with probability at least $1 - 2\exp(-p)$, the following holds for $np > 51 \max\{\chi, \chi^2\}$

$$\sup_{w\in\mathbb{B}^\delta(\tilde{w})} |\frac{1}{m} \sum_{i=1}^{m} g(\langle x_i, w\rangle) - \mathbb{E}[g(\langle x, w\rangle)]| \leq c(\kappa_g + \frac{\kappa_x}{\tilde{u}}) \sqrt{\frac{p \log m}{m}},$$

where $\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta^2 G^2 \tilde{\mu}^2}$. $c$ is some absolute constant.

**Lemma 8.** Let $\mathbb{B}^\delta(\tilde{w})$ be the ball centered at $\tilde{w}$ and with radius $\delta$ (*i.e.*, $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w-\tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2\cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d sub-Gaussian random vectors with covariance matrix $\Sigma$. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is uniformly bounded by $L$ and Lipschitz continuous with $G$, the following holds with probability at least $1 - \exp(-p)$

$$\sup_{w\in\mathbb{B}^\delta(\tilde{w})} |\frac{1}{m}\sum_{i=1}^m g(\langle x_i, w\rangle) - \mathbb{E}[g(\langle x, w\rangle)]| \leq 2\{G(\|\tilde{w}\|_2+\delta)\|\Sigma\|_2 + L\}\sqrt{\frac{p}{m}}.$$

The following lemma shows that the private estimator $\hat{w}^{ols}$ is close to the unperturbed one.

**Lemma 9.** Let $X = [x_1^T; x_2^T; \cdots; x_n^T] \in \mathbb{R}^{n\times d}$ be a matrix such that $X^TX$ is invertible, and $x_1, \cdots, x_n$ are realizations of a sub-Gaussian random variable $x$ which satisfies the condition of $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x = O(1)$ and $\Sigma = \mathbb{E}[xx^T]$ is the the population covariance matrix. Let $\tilde{w}^{ols} = (X^TX)^{-1}X^Ty$ denote the empirical linear regression estimator. Then, for sufficiently large $n \geq \Omega(\frac{\kappa_x^4\|\Sigma\|_2^2 pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda^2_{\min}(\Sigma)})$, the following holds with probability at least $1 - \exp(-\Omega(p)) - \xi$,

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\frac{pr^2(1 + r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2 n\lambda^2_{\min}(\Sigma)}\big), \tag{10}$$

where $r = r$ if $x_i$ is sampled from some bounded distribution.

*Proof of Lemma 9.* It is obvious that $\widehat{X^TX} = X^TX + E_1$, where $E_1$ is a symmetric Gaussian matrix with each entry sampled from $\mathcal{N}(0, \sigma_1^2)$ and $\sigma_1^2 = O(\frac{nr^4\log\frac{1}{\delta}}{\epsilon^2})$. $\widehat{X^Ty} = X^Ty + E_2$, where $E_2$ is a Gaussian vector sampled from $\mathcal{N}(0, \sigma_2^2 I_p)$ and $\sigma_2^2 = O(\frac{nr^2\log\frac{1}{\delta}}{\epsilon^2})$.

We first show that $\widehat{X^TX}$ is invertible with high probability under our assumption.

It is sufficient to show that $X^TX + E_1 \succ \frac{X^TX}{2}$, *i.e.*, $\|E_1\|_2 \leq \frac{\lambda_{\min}(X^TX)}{2}$. By Lemma 6, we can see that with probability $1 - \exp(-\Omega(p))$,

$$\|E_1\|_2 \leq O(\frac{r^2\sqrt{pn\log\frac{1}{\delta}}}{\epsilon}).$$

Also, by Lemma 5 and Lemma 3 we know that with probability at least $1 - \exp(-\Omega(p))$,

$$\lambda_{\min}(X^TX) \geq n\lambda_{\min}(\Sigma) - O(\kappa_x^2\|\Sigma\|_2\sqrt{pn}).$$

Thus, it is sufficient to show that $n\lambda_{\min}(\Sigma) \geq O(\frac{\kappa_x^2\|\Sigma\|_2 r^2\sqrt{pn\log\frac{1}{\delta}}}{\epsilon})$, which is true under the assumption of $n \geq \Omega(\frac{\kappa_x^4\|\Sigma\|_2^2 pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda^2_{\min}(\Sigma)})$. Thus, with probability at least $1 - \exp(-\Omega(p))$, it is invertible. In the following we will always assume that this event holds.

By direct calculation we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2 = -(X^TX + E_1)^{-1}E_1\tilde{w}^{ols} + (X^TX + E_1)^{-1}E_2.$$

Thus, by Cauchy-Schwartz inequality we get

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\|E_1\tilde{w}^{ols}\|^2_{(X^TX+E_1)^{-2}} + \|E_2\|^2_{(X^TX+E_1)^{-2}}\big).$$

Since we already assume that $X^TX + E_1 \succ \frac{X^TX}{2}$, by Lemma 4 we can obtain the following with probability at least $1 - \xi$

$$\|E_1\tilde{w}^{ols}\|^2_{(X^TX+E_1)^{-2}} \leq O\big(\frac{nr^4\log\frac{1}{\delta}}{\epsilon^2}\|\tilde{w}^{ols}\|_2^2\text{Tr}((X^TX)^{-2})\log\frac{4p^2}{\xi}\big)$$

$$\|E_2\|^2_{(X^TX+E_1)^{-2}} \leq O\big(\frac{nr^2\log\frac{1}{\delta}}{\epsilon^2}\text{Tr}((X^TX)^{-2})\frac{4p}{\xi}\big).$$

Thus, we have

$$\|\hat{w}^{ols}-\tilde{w}^{ols}\|_2^2 \leq C_1 n\cdot\frac{r^2(1+r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2}\text{Tr}((X^TX)^{-2}).$$

For the term of $\text{Tr}((X^TX)^{-2})$, we get

$$\text{Tr}((X^TX)^{-2})\leq(\text{Tr}((X^TX)^{-1}))^2\leq p\|(X^TX)^{-2}\|_2^2=\frac{p}{\lambda^2_{\min}(X^TX)}\leq O(\frac{p}{n^2\lambda^2_{\min}(\Sigma)}),$$

where the last inequality is due to the fact that $\lambda_{\min}(X^TX)\geq n\lambda_{\min}(\Sigma)-O(\kappa_x^2\|\Sigma\|_2\sqrt{pn})\geq\frac{1}{2}n\lambda_{\min}(\Sigma)$ (by the assumption on $n$). This completes the proof. $\square$

Let $w^{ols}=(\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$ denote the population linear regression estimator. The following lemma bounds the estimation error between $\tilde{w}^{ols}$ and $w^{ols}$. The proof could be found in [16] or [11].

**Lemma 10** (Prop. 7 in [16])**.** Assume that $\mathbb{E}[x_i]=0$, $\mathbb{E}[x_ix_i^T]=\Sigma$, and $\Sigma^{-\frac{1}{2}}x_i$ and $y_i$ are sub-Gaussian with norms $\kappa_x$ and $\gamma$, respectively. If $n\geq\Omega(\kappa_x\gamma p)$, the following holds

$$\|\tilde{w}^{ols}-w^{ols}\|_2\leq O\big(\gamma\kappa_x\sqrt{\frac{p}{n\lambda_{\min}(\Sigma)}}\big),$$

with probability at least $1-3\exp(-p)$.

# B Proofs of LDP

The LDP proof of Algorithm 1 follows from Gaussian mechanism and the composition property of DP.

For Algorithm 2, it is $(\epsilon,\delta)$-LDP due to the $\ell_2$-norm bound on $\|x_iy_i\|_2=\|x_i\|_2\|f(\langle x,w^*\rangle)+\sigma_i\|_2\leq\|x_i\|_2(L\|x\|_2+|f(0)|+C)$, where the last inequality is due to the fact that $f'$ is $L$-bounded and $\|w^*\|_2\leq 1$. That is, $|f(\langle x,w^*\rangle)-f(0)|\leq L|\langle x,w^*\rangle-0|\leq L\|x\|_2\|w^*\|_2$.

# C Proofs and Comments in Section 4

Since Theorem 3 is the most complicated one, we will first prove it and then Theorem 2.

## C.1 Proof of Theorem 3

Since $r=O(1)$ (by assumption), combining this with Lemmas 9 and 10, we have that with probability at least $1-\exp(-\Omega(p))-\xi$ and under the assumption on $n$, there is a constant $C_3>0$ such that

$$\|\hat{w}^{ols}-w^{ols}\|_2\leq C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda^{1/2}_{\min}(\Sigma)\min\{\lambda^{1/2}_{\min}(\Sigma),1\}}. \tag{11}$$

**Lemma 11.** Let $\Phi^{(2)}$ be a function that is Lipschitz continuous with constant $G$, and $f:\mathbb{R}\times\mathbb{R}^p\mapsto\mathbb{R}$ be another function such that $f(c,w)=c\mathbb{E}[\Phi^{(2)}(\langle x,w\rangle c)]$ and its empirical one is

$$\hat{f}(c,w)=\frac{c}{m}\sum_{j=1}^m\Phi^{(2)}(\langle x,w\rangle c).$$

Let $\mathbb{B}^\delta(\bar{w}^{ols})=\{w:\|w-\bar{w}^{ols}\|_2\leq\delta\}$, where $\bar{w}^{ols}=\Sigma^{\frac{1}{2}}w^{ols}$. Under the assumptions in Lemma 9 and Eq. (11), if further assume that $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2}\leq\kappa_x$, $\sup_{w\in\mathbb{B}^\delta(\bar{w}^{ols})}\|\Phi^{(2)}(\langle x,w\rangle)\|_{\psi_2}\leq\kappa_g$, and there exist $\bar{c}>0$ and $\tau>0$ such that $f(\bar{c},w^{ols})\geq 1+\tau$, then there is $\bar{c}_\Phi\in(0,\bar{c})$ such that $1=f(\bar{c}_\Phi,w^{ols})$. Also, for sufficiently large $n$ and $m$ such that

$$m\geq\Omega\big((\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2\max\{p\log m\tau^{-2},\frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\}\big), \tag{12}$$

$$n\geq\Omega(\kappa_x^4G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big), \tag{13}$$

with probability at least $1-2\exp(-p)$, there exists a $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in the absolute value (*i.e.,* does not change sign) by $M>0$ in the interval $c\in[0,\bar{c}]$, then the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(M^{-1}\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \tag{14}$$

**Proof of Lemma 11.** We divide the proof into three parts.

**Part 1: Existence of $\bar{c}_\Phi$:** From the definition, we know that $f(0, w^{ols}) = 0$ and $f(\bar{c}, w^{ols}) > 1$. Since $f$ is continuous, we known that there exists a constant $\bar{c}_\Phi \in (0,\bar{c})$ which satisfying $f(\bar{c}_\Phi, w^{ols}) = 0$.

**Part 2: Existence of $\hat{c}_\Phi$:** For simplicity, we use the following notations.

$$\delta = C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}, \delta' = \frac{\|\Sigma\|_2^{\frac{1}{2}}\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}, \tag{15}$$

where $C_3$ is the one in (11). Thus, $\|\Sigma^{\frac{1}{2}}\hat{w}^{ols} - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'$.

Now consider the term of $|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})|$ for $c\in[0,\bar{c}]$. We have

$$\sup_{c\in[0,\bar{c}]}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \leq \sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,w) - f(c,w)|, \tag{16}$$

where $\mathbb{B}_\Sigma^{\delta'}(w^{ols}) = \{w : \|\Sigma^{\frac{1}{2}}w - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'\}$.

Note that for any $x$, we have $\langle x, w\rangle = \langle v, \Sigma^{\frac{1}{2}}w\rangle$, where $v = \Sigma^{-\frac{1}{2}}x$ follows an isotropic sub-Gaussian distribution. Also, by definition we know that $w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})$ is equivalent to $\Sigma^{\frac{1}{2}}w \in \mathbb{B}^{\delta'}(\bar{w}^{ols})$. Thus, we have

$$\begin{aligned}
&\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \\
&\leq \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\frac{1}{m}\sum_{j=1}^m \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{\Sigma^{\frac{1}{2}}w\in\mathbb{B}^{\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m \Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)|. 
\end{aligned} \tag{17}$$

By Lemma 7, we know that when $mp \geq 51\max\{\chi, \chi^{-1}\}$, where

$$\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta'^2G^2\tilde{\mu}^2} = \Theta\big(\frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\big),$$

the following holds with probability at least $1-2\exp(-p)$

$$\sup_{w'\in\mathbb{B}^{\bar{c}\delta}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m \Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \leq O((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}). \tag{18}$$

By the Lipschitz property of $\Phi^{(2)}$, we have that for any $w_1$ and $w_2$,

$$\begin{aligned}
\sup_{c\in[0,\bar{c}]}|f(c,w_1) - f(c,w_2)| &\leq G\bar{c}^2\mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(w_1 - w_2)\rangle] \\
&\leq \kappa_x G\bar{c}^2\|\Sigma^{\frac{1}{2}}(w_1 - w_2)\|_2.
\end{aligned} \tag{19}$$

Taking $w_1 = \hat{w}^{ols}$ and $w_2 = w^{ols}$, we have

$$\sup_{c\in[0,\bar{c}]} |f(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\big(\kappa_x G\bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}\big).$$

Combining this with (17), (18), (19), and taking $\delta$ as in (15), we get

$$\sup_{c\in[0,\bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\big(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big). \tag{20}$$

Let $B$ denote the RHS of (20). If $c = \bar{c}$, we have $\hat{f}(c, \hat{w}^{ols}) \geq 1 + \tau - B$. Thus, if $B \leq \tau$, there must exist a $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$.

To ensure that $B \leq \tau$ holds, it is sufficient to have

$$O(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}) \leq \frac{\tau}{2}$$

and

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}) \leq \frac{\tau}{2}.$$

This means that

$$m \geq \Omega\big(\bar{c}^2(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 p\log m\tau^{-2}\big),$$
$$n \geq \Omega(\kappa_x^4 G^2\bar{c}^4\|\Sigma\|_2 \frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}),$$

which are assumed in the lemma.

**Part 3: Estimation Error:** So far, we know that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = f(\bar{c}_\Phi, w^{ols}) = 1$ with high probability. By (16), (17) and (18), we have

$$|1 - f(\hat{c}_\Phi, \hat{w}^{ols})| = |\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) - f(\hat{c}_\Phi, \hat{w}^{ols})| \leq O(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}).$$

By the same argument for (20), we have

$$|f(\hat{c}_\Phi, \hat{w}^{ols}) - f(\hat{c}_\Phi, w^{ols})| \leq G\kappa_x\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}.$$

Thus, using Taylor expansion on $f(c, w^{ols})$ around $c_\Phi$ and by the assumption of the bounded derivative of $f$, we have

$$\begin{aligned}
M|\hat{c}_\Phi - \bar{c}_\Phi| &\leq |f(\hat{c}_\Phi, w^{ols}) - f(\bar{c}_\Phi, w^{ols})| \\
&\leq |f(\hat{c}_\Phi, w^{ols}) - f(\hat{c}_\Phi, \hat{w}^{ols})| + |f(\hat{c}_\Phi, \hat{w}^{ols}) - 1| \\
&\leq O\big(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big).
\end{aligned}$$

□

Next, we prove our main theorem.

**Proof of Theorem 3.** By definition, we have

$$\begin{aligned}
\|\hat{w}^{glm} - w^*\|_\infty &\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\
&\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty.
\end{aligned} \tag{21}$$

We first bound the term of $|\bar{c}_\Phi - c_\Phi|$. Since $\bar{c}_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle \bar{c}_\Phi)] = 1$ and $c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle)] = 1$ (by definition), we get

$$
\begin{aligned}
|f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| &= |c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle)] - f(c_\Phi, w^{ols})| \\
&\leq c_\Phi |\mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle) - \Phi^{(2)}(\langle x, w^{ols}\rangle c_\Phi)] \\
&\leq c_\Phi G |\mathbb{E}[\langle x, (w^* - c_\Phi w^{ols})\rangle] \\
&\leq c_\Phi G \|(w^* - c_\Phi w^{ols})\|_\infty \mathbb{E}\|x\|_1 \\
&\leq c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty,
\end{aligned}
$$

where the last inequality is due to the assumption that $\|x\|_1 \leq r$.

Thus, by the assumption of the bounded deviation of $f(c, w^{ols})$ on $[0, \max\{\bar{c}, c_\Phi\}]$, we have

$$
M|\bar{c}_\Phi - c_\Phi| \leq |f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| \leq c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty.
$$

By Lemma 2, we have

$$
|\bar{c}_\Phi - c_\Phi| \leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}. \tag{22}
$$

Thus, the second term of (21) is bounded by

$$
\begin{aligned}
&\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty \leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}} \|w^{ols}\|_\infty \\
&\leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^3}{\sqrt{p}} \big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\big) \\
&= O\big(M^{-1} r^3 \kappa_x^6 G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, c_\Phi\}\big),
\end{aligned} \tag{23}
$$

where the last inequality is due to Lemma 2.

By Lemma 2, the third term of (21) is bounded by $16 c_\Phi G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}$.

For the first term of (21), by (11) and Lemma 11 we have

$$
\begin{aligned}
&\|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty \leq |\hat{c}_\Phi| \cdot \|\hat{w}^{ols} - w^{ols}\|_\infty + |\hat{c}_\Phi - \bar{c}_\Phi| \cdot \|w^{ols}\|_\infty \\
&\leq O\big(\bar{c} \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&+ \|w^{ols}\|_\infty (M^{-1} \bar{c} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}} + M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}})\big).
\end{aligned} \tag{24}
$$

For the first term of (24), we have

$$
\begin{aligned}
&\bar{c} \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \leq \bar{c} \frac{\kappa_x p r^2 \|w^{ols}\|_\infty \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&\leq \bar{c} \frac{\kappa_x p r^2 \|w^*\|_\infty \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\big) \\
&= O\big(\bar{c} \frac{p \kappa_x^4 \sqrt{\rho_2} \rho_\infty G r^3 \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} \sqrt{\log\frac{1}{\delta} \log\frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}\big).
\end{aligned} \tag{25}
$$

For the second term of (24), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\\
&\leq \bar{c}\|w^*\|_\infty(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})\\
&\leq O\big(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big). \qquad (26)
\end{aligned}$$

For the third term of (24), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}})\\
&\leq M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{pr^2\|w^*\|_\infty^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})^2\\
&\leq O\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \qquad (27)
\end{aligned}$$

Thus, the first term of (21) is bounded by (since $m\geq\Omega(n)$)

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty\leq O\big(\bar{c}\frac{p\kappa_x^4\sqrt{\rho_2}\rho_\infty Gr^3\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}\\
&+Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}+\\
&M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&=O\big(M^{-1}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log m\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big).
\end{aligned}$$

Putting all the bounds together, we have

$$\begin{aligned}
&\|\hat{w}^{glm}-w^*\|_\infty\leq\tilde{O}\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&+M^{-1}r^3\kappa_x^6c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}+\\
&Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big). \qquad (28)
\end{aligned}$$

Next, we bound the probability. We assume that Lemma 9, 10 and 11 hold with probability at least $1-\exp(-\Omega(p))-\rho$. They hold when

$$m\geq\Omega\big((\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2\max\{p\log m\tau^{-2},\frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}\}\big), \qquad (29)$$

$$n\geq\Omega(\max\{\kappa_x^4G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}},\frac{\kappa_x^4\|\Sigma\|_2^2pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda_{\min}^2(\Sigma)}\}\big). \qquad (30)$$

Since $\|w^{ols}\|_2 \leq \sqrt{p}\|w^*\|_\infty(\frac{1}{c_\Phi} + 16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})$, it suffices for $n$

$$n \geq \Omega\big(G^4\bar{c}^4\|\Sigma\|_2^2\frac{p^2r^6\kappa_x^{10}\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\max\{1, \frac{1}{c_\Phi}\}^2\big). \tag{31}$$

□

### C.2 Proof of Theorem 2

**Lemma 12.** Let $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$ be defined the same as in Lemma 11. If further assume that $|\Phi^{(2)}(\cdot)| \leq L$ for some constant $L > 0$ and is Lipschitz continuous with constant $G$, then, under the assumptions in Lemma 9 and (11), with probability at least $1 - 4\exp(-p)$ there exists a constant $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in absolute value (*i.e.,* does not change the sign) by $M > 0$ in the interval $c \in [0, \bar{c}]$, then with probability at least $1 - 4\exp(-p)$, the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(\frac{M^{-1}GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{32}$$

for sufficiently large $m, n$ such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big) \tag{33}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \tag{34}$$

*Proof of Lemma 12 .* The main idea of this proof is almost the same as the one for Lemma 11. The only difference is that instead of using Lemma 7 to get (18), we use here Lemma 8 to obtain the following with probability at least $1 - \exp(-p)$

$$\begin{aligned}
&\sup_{w' \in \mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^{m}\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \\
&\leq O\big((G(\|\bar{w}^{ols}\|_2 + \bar{c}\delta')\|I\|_2 + L)\sqrt{\frac{p}{m}} \\
&\leq O\big((G\|\Sigma\|_2^{\frac{1}{2}}(\|w^{ols}\|_2 + \bar{c}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}) + L)\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{35}$$

Thus, by (17), (19) and (35), we have

$$\begin{aligned}
\sup_{c \in [0, \bar{c}]}|\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| &\leq O\big(G\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}} + \\
&\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}} + L\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{36}$$

Let D denote the RHS of (36), we have

$$\hat{f}(\bar{c}, \hat{w}^{ols}) \geq 1 + \tau - D.$$

It is sufficient to show that $\tau > D$, which holds when

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}) \leq \frac{\tau}{2}$$

and

$$O(\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}L\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}}) \leq \frac{\tau}{2}.$$

That is,

$$n \geq \Omega\big(\frac{G^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4 pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{37}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{38}$$

Then, there exists $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols})=1$. We can easily get

$$\begin{aligned}
M|\hat{c}_\Phi-\bar{c}_\Phi| &\leq |f(\hat{c}_\Phi,w^{ols})-f(\bar{c}_\Phi,w^{ols})| \\
&\leq O\big(\frac{G\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}} \\
&+\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) &(39)\\
&\leq O\big(\frac{GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big). &(40)
\end{aligned}$$

□

**Proof of Theorem 2 .** The proof is almost the same as the one for Theorem 3. By definition, we have

$$\begin{aligned}
\|\hat{w}^{glm}-w^*\|_\infty &\leq \|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-w^*\|_\infty \\
&\leq \|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty+\|c_\Phi w^{ols}-w^*\|_\infty. &(41)
\end{aligned}$$

The second term of (41) is bounded by

$$\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty \leq O\big(M^{-1}r^2\kappa_x^7c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\big). \tag{42}$$

By Lemma 2, the third term of (41) is bounded by $16c_\Phi Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}}$. The first term is bounded by

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty \leq \\
&O\big(\frac{M^{-1}G^3L\bar{c}^2\kappa_x^8r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\times\max\{\frac{1}{c_\Phi},1\}^2 \\
&\quad+\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}\times\max\{\frac{1}{c_\Phi},1\}^2\big). &(43)
\end{aligned}$$

Thus, in total we have

$$\begin{aligned}
\|\hat{w}^{glm}-w^*\|_\infty \leq O\big(&\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}\times\max\{\frac{1}{c_\Phi},1\}^2 \\
&+\frac{G^3L\bar{c}^2\kappa_x^6r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\max\{\frac{1}{c_\Phi},1\}^2 \\
&+M^{-1}r^2\kappa_x^7c_\Phi G^3\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\big). &(44)
\end{aligned}$$

The probability of success is at least $1-\exp(-\Omega(p))-\xi$. The sample complexity should satisfy

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}G^2r^2\kappa_x^6\rho_2\rho_\infty^2p^2\tau^{-2}\max\{1,\frac{1}{c_\Phi}\}^2\big) \tag{45}$$

$$n \geq \Omega\big(\frac{\rho_2\rho_\infty^2G^4\tau^{-2}\bar{c}^4\|\Sigma\|_2^2\kappa_x^{10}p^2\|w^*\|_\infty^2r^6\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^3}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \tag{46}$$

□

# D Proofs in Section 5

## D.1 Proof of Theorem 4

The idea of the proof follows the one in [16].

**Proof of Theorem 4.** By assumption, we have

$$\mathbb{E}[xy] = \mathbb{E}[xf(\langle x, w^* \rangle)] = \Sigma^{\frac{1}{2}} \mathbb{E}[vf(\langle v, \hat{w}^* \rangle)],$$

where $\hat{w}^* = \Sigma^{\frac{1}{2}} w^*$. Now, consider each coordinate $j \in [p]$ for the term $\mathbb{E}[vf(\langle v, \hat{w}^* \rangle)]$. Let $v_j^*$ denote the zero-bias transformation of $v_j$ conditioned on $V_j = \langle v, \hat{w}^* \rangle - v_j \hat{w}_j^*$. Then, we have

$$\begin{aligned}
\mathbb{E}[v_j f(\langle v, \hat{w}^* \rangle)] &= \mathbb{E}\mathbb{E}[v_j f(v_j \hat{w}_j^* + V_j)|V_j] \\
&= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'(v_j^* \hat{w}_j^* + V_j)|V_j] \\
&= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^* \rangle)|V_j] \\
&= \hat{w}_j^* \mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^* \rangle)].
\end{aligned}$$

Thus, we have $w^{ols} = \Sigma^{-\frac{1}{2}} D \Sigma^{\frac{1}{2}} w^*$, where $D$ is a diagonal matrix whose $i$-th entry is $\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^* \rangle)]$.

By the Lipschitz condition, we have

$$|\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^* \rangle)] - \mathbb{E}[f'(\langle v, \hat{w}^* \rangle)]| \leq G|\hat{w}_j^*|\mathbb{E}|(v_j^* - v_j)|.$$

By the same argument given in [16], we have

$$\mathbb{E}|(v_j^* - v_j)| \leq 1.5\mathbb{E}[|v_j|^3].$$

Using the bound of the third moment induced by the sub-Gaussian norm, we have

$$L|\hat{w}_j^*|\mathbb{E}|(v_j^* - v_j)| \leq 8G\kappa_x^3 \max_{j \in [p]} |\hat{w}_j^*| \leq 8G\kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

Thus, we get

$$\max_{j \in [d]} |D_{jj} - \frac{1}{c_f}| \leq 8G\kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

This means that

$$\begin{aligned}
\|w^{ols} - \frac{1}{c_f} w^*\|_\infty &= \|\Sigma^{-\frac{1}{2}}(D - \frac{1}{c_f} I)\Sigma^{\frac{1}{2}} w^*\|_\infty \\
&\leq \max_{j \in [p]} |D_{jj} - \frac{1}{c_f}| \|\Sigma^{-\frac{1}{2}}\|_\infty \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty \\
&\leq 8L\kappa_x^3 \rho_\infty L \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty^2.
\end{aligned}$$

Due to the diagonal dominance property we have

$$\|\Sigma^{\frac{1}{2}}\|_\infty = \max_i \sum_{j=1}^p |\Sigma_{ij}^{\frac{1}{2}}| \leq 2 \max \Sigma_{ii}^{\frac{1}{2}} \leq 2\|\Sigma\|_2^{\frac{1}{2}}.$$

Since we have $\|x\|_2 \leq r$, we write

$$r^2 \geq \mathbb{E}[\|x\|_2^2] = \text{Trace}(\Sigma) \geq p\|\Sigma \geq \frac{p\|\Sigma\|_2}{\rho_2}.$$

Thus we have $\|\Sigma^{\frac{1}{2}}\|_\infty \leq 2r\sqrt{\frac{\rho_2}{p}}$. □

### D.2 Proof of Theorem 5

By the same argument in the proof of Lemma 9, we can show that when $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$, with probability at least $1 - \exp(-\Omega(p)) - \xi$, the following holds

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\frac{pC^2 r^2 (L^2 r^2 + C^2 + r^2 \|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}\big). \tag{47}$$

Thus, by Lemma 10 we have

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq O\big(\frac{CL\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big). \tag{48}$$

In the following, we will always assume that (48) holds. By the same argument given in Lemma 12, we have the following Lemma, which can be proved in the same way as Lemma 12.

**Lemma 13.** Let $f'$ be a function that is Lipschitz continuous with constant $G$ and $|f'(\cdot)| \leq L$, and $g : \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$ be another function such that $g(c, w) = c\mathbb{E}[f'(\langle x, w \rangle c)]$ and its empirical one is

$$\hat{g}(c, w) = \frac{c}{m} \sum_{j=1}^{m} f'(\langle x, w \rangle c).$$

Let $\mathbb{B}^\delta(\bar{w}^{ols}) = \{w : \|w - \bar{w}^{ols}\|_2 \leq \delta\}$, where $\bar{w}^{ols} = \Sigma^{\frac{1}{2}} w^{ols}$. Then, under the assumptions in Lemma 9 and Eq. (48), with probability at least $1 - 4\exp(-p)$, there exists a constant $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{g}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto g(c, w^{ols})$ is bounded below in absolute value (*i.e.,* does not change the sign) by $M > 0$ in the interval of $c \in [0, \bar{c}]$, then with probability at least $1 - 4\exp(-p)$, the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(\frac{M^{-1} CGL\bar{c}^2 r^2 \|\Sigma\|_2^{\frac{1}{2}} \sqrt{p} \|w^{ols}\|_2 \log \frac{1}{\delta} \log \frac{p}{\xi^2}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} + M^{-1} LG \|\Sigma\|_2^{\frac{1}{2}} \|w^{ols}\|_2 \sqrt{\frac{p}{m}}\big) \tag{49}$$

for sufficiently large $m, n$ such that

$$n \geq \Omega\big(\frac{LG^2 \tau^{-2} \bar{c}^4 \|\Sigma\|_2 \kappa_x^4 p r^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}\big) \tag{50}$$

$$m \geq \Omega\big(G^2 L^2 \|\Sigma\|_2 \|w^{ols}\|_2^2 p \tau^{-2}\big). \tag{51}$$

where $r = \max_{i \in [n]} \|x_i\|_2$.

## E Details of the experiments

We conduct experiment for GLM with logistic loss on the Covertype dataset [13]. Before running our algorithm, we first normalize the data and remove some co-related features. After the pre-processing, the dataset contains 581012 samples and 44 features. There are seven possible values for the label. Since multinomial logistic regression can not be regarded as a Generalized Linear Model, we consider a weaker test, which is to classify whether the label is Lodgepole Pine (type 2) or not. The chosen algorithm is still binary logistic regression. We divide the data into training and testing, where $n_{\text{training}} = 406708$ and $n_{\text{testing}} = 174304$ and randomly choose the sample size $n \in 10^4 \cdot \{1, 2, 3, ..., 39\}$ from the training data and use the same amount of public data. Regarding the privacy parameter, we take $\delta = \frac{1}{n}$ and let $\epsilon$ take value from $\{20, 10, 5\}$. We measure the performance by the prediction accuracy. For each combination of $\epsilon$ and $n$, the experiment is repeated 1000 times. We observe that when $\epsilon$ takes a reasonable value, the performance is approaching to the non-private case, provided that the size of private dataset is large enough. Thus, our algorithm is practical and is comparable to the non-private one.

We also conduct experiment for GLM with logistic loss on the SUSY dataset [1]. The task is to classify whether the class label is signal or background. After the pre-processing and sampling, the dataset contains 500000 samples and 18 features. Then we divide the data into training and testing, where $n_{\text{training}} = 350000$ and $n_{\text{testing}} = 150000$ and randomly choose the sample size $n \in 10^4 \cdot \{1, 3, \cdots, 33\}$ from the training data and use the same amount of public data. Regarding the privacy parameter, we take $\delta = \frac{1}{n}$ and let $\epsilon$ take value from $\{20, 10, 5\}$. We measure the performance by the prediction accuracy. For each combination of $\epsilon$ and $n$, the experiment is repeated 1000 times.

### E.1 The effect of public unlabeled data

We use similar setting as our synthetic experiments in Section 6.1. For GLM we consider the problem of binary logistic loss *i.e.,* $\Phi(\langle x, w\rangle) = \ln(1 + \exp(\langle x, w\rangle))$ in (1) while for non-linear regression we will set $f(x) = \frac{1}{3}x^3$ in (7). We compare relative error $\frac{\|\hat{w} - w^*\|_\infty}{\|w^*\|_\infty}$ with respect to different privacy parameters $\epsilon \in \{10, 5, 3\}$ with $\delta = \frac{1}{n}$. In these experiments, we fix dimensionality $p = 10$ and the population parameter $w^* = (1, 1, ..., 1)/\sqrt{p}$. We also fix the private sample size $n = 200000$ and the public data size is chosen from the set $10^4 \cdot \{2, 4, ..., 16\}$. We assume that the same amount of public unlabeled data is available. The features are generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$ and the label is generated according to the logistic model or the model (7). In non-linear regression model, $\sigma$ is bounded by $C = 0.001$. The results are shown in Figure 4a and 4b. We can see that sometimes there is no need to use as large amount of public data as our theoretical result requires to guarantee a good performance, as is shown by Figure 4a.

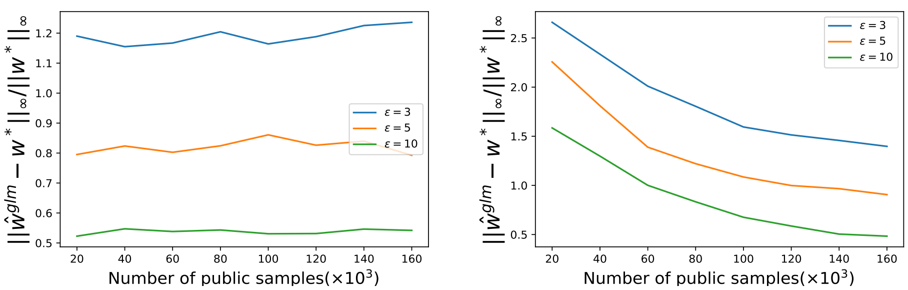



Figure 4: The effect of the number of public unlabeled samples. The left plot shows the relative error of GLM with logistic loss. The right one shows the relative error of cubic regression.